# METHOD FOR ANALYSING SYNTAX BASED ON MOBILE CONFIGURATION CONCEPT AND METHOD FOR SEARCHING NATURAL LANGUAGE USING THE SAME

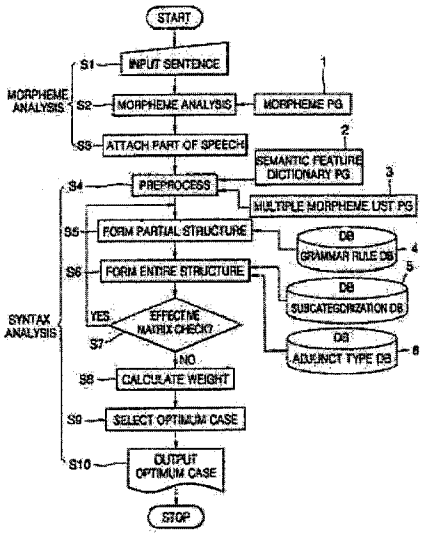
**Publication number:** KR20030044949 (A)
**Publication date:** 2003-06-09
**Inventor(s):** WOO SOON JO [KR]
**Applicant(s):** WOO SOON JO [KR]
**Classification:**
**- international:** ***G06F17/27;*** **G06F17/27;** (IPC1-7): G06F17/27
**- European:** G06F17/27G
**Application number:** KR20030025995 20030424
**Priority number(s):** KR20030025995 20030424

**Also published as:**
WO2004095310 (A1)
US2007010990 (A1)
JP2007317211 (A)
JP2006524372 (T)
HK1092242 (A1)
EP1616270 (A1)
CN1777888 (A)
CN100378724 (C)
CA2523140 (A1)

<< less

Abstract of **KR 20030044949 (A)**

PURPOSE: A method for analyzing a syntax based on a mobile configuration concept and a method for searching a natural language using the same are provided to supply core technique data necessary for developing various and useful tools and supply a universality and a reliability based on a strict linguistic result and to improve a performance continuously and rapidly by increasing an independence between a language knowledge and an analysis engine. CONSTITUTION: A morpheme dictionary program for judging a postposition, a verbal word end as an independent basic element and storing a characteristic of a grammatical function of a word end as a morpheme dictionary is constructed. A grammar rule database is constructed. If a sentence is inputted(S1), a morpheme is analyzed by the morpheme dictionary program(S2) and a tag is attached to a part-of-speech(S3).; If a morpheme having a tagged part-of-speech is inputted, it is judged whether a sentence structure of a multi-morpheme type exists. If a sentence structure of a multi-morpheme type exists, the morpheme is converted into a multi-morpheme type(S4). If a meaning disposition part-of-speech attached morpheme is inputted, individual morphemes are processed. It is judged whether a local structure exists based on a grammatical rule. A local structure is formed. In addition, a reflexive local structure is formed by referring to a succession process object(S5). The total structure is formed according to categories of sentence structure and formula patterns based on the lower category database and an attached word pattern database (S6). It is judged whether a valid matrix of other pattern is checked(S7) and a partial structure of the next matrix is formed(S5).; A weight value is calculated based on importance of each structure based on a position and a character of a sentence structure(S8), the optimum example is selected(S9), and the selected optimum example is output(S10).



(19)대한민국특허청(KR)
(12) 공개특허공보(A)

(51) 。Int. Cl. 7
G06F 17/27

(11) 공개번호 특2003-0044949
(43) 공개일자 2003년06월09일

(21) 출원번호 10-2003-0025995
(22) 출원일자 2003년04월24일

(71) 출원인 우순조
경기도 성남시 분당구 서현동 96 시범우성아파트 228동605호

(72) 발명자 우순조
경기도 성남시 분당구 서현동 96 시범우성아파트 228동605호

(74) 대리인 이영필
이해영

심사청구 : 있음

# (54) 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를이용한 자연어 검색 방법

요약

본 발명은 모빌적 형상 개념을 기초로 한 구문 분석방법 및 자연어 검색 방법에 관한 것으로서, 입력된 문장의 형태소를 분석하는 형태소 사전 프로그램, 조사와 어미를 모두 통사의 단위로 취급하는 표지이론에 입각하여 용언 어미의 통사적 지위를 인정하고, 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어간 및 어미 등 중심어가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 구축하고, 분석할 문장이 입력되면, 상기 형태소 사전 프로그램에 의해 어절별 형태소 내역을 분석하고, 어절별 형태소 분석자료 중 해당 입력 자료에 적합한 형태소의 분석 사례를 선택하여 전처리를 실시하는 형태소 분석단계; 및 분석된 형태소들을 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 상기 하위범주화 데이터베이스를 이용하여 전체적인 구조를 확립하며, 각 구조의 가중치를 계산하여 가장 적합한 최적례를 확정하여 출력하는 구문 분석단계;를 포함하여 이루어지는 것을 특징으로 하기 때문에 어떠한 어순 도치형 구문도 분석이 용이하여 별도의 어려운 분석 장치 없이 빠르게 처리할 수 있으며, 문장을 구성하는 표현들 사이의 문법적 관계를 정확하게 포착해 낼 수 있으며, 그 결과 사람이 판단하는 바와 같은 방식으로 사용자가 요구하는 정보를 검색해서 정확한 정보를 제공할 수 있는 효과를 갖는다.

대표도

도 1

명세서

도면의 간단한 설명

도 1은 본 발명의 바람직한 일 실시예에 따른 모빌적 형상 개념을 기초로 한 구문 분석방법을 나타내는 순서도이다.

도 2는 도 1의 전처리 단계의 일례를 보다 상세하게 나타내는 순서도이다.

도 3은 도 1의 부분 구조 형성 단계의 일례를 보다 상세하게 나타내는 순서도이다.

도 4는 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 결과 화면의 일례를 나타내는 도면이다.

도 5는 본 발명의 바람직한 일 실시예에 따른 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법을 나타내는 순서도이다.

도 6은 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 시스템의 질의어(검색어) 입력 화면 및 결과 화면의 일례를 나타내는 도면이다.

도 7 내지 도 11은 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색방법의 내부 데이터베이스의 일례를 단계적으로 나타내는 도 면이다.

도 12는 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법의 프린트 화면의 일례를 나타내는 도면이다.

## 발명의 상세한 설명

### 발명의 목적

#### 발명이 속하는 기술 및 그 분야의 종래기술

본 발명은 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법에 관한 것으로서, 더욱 상세하게는 하위범주화 정보에 미리 규정되어 있는 문법적 기능 정보가 직접 구성성분에 부여되어 자유어순에 능동적으로 대처할 수 있게 하는 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법에 관한 것이다.

구문 분석이란 한마디로 컴퓨터를 이용하여 자연 언어의 통사적 구조를 분석하게 하는 것이다. 즉, 이러한 구문 분석을 위해서는 컴퓨터에 자연 언어에 대한 지식을 전달 및 구현하는 것이 반드시 필요하다.

다시 말해, 자연 언어 처리 방법의 개발은 컴퓨터에 말을 가르치는 것이라 요약할 수 있고 이러한 기존의 구문 분석은 확률기반적방식을 사용하고 있다.

여기서, 기존의 확률기반적 구문 분석방법이란, 대량의 코퍼스(corpus)를 구축하여 거기로부터 국부 구조 및 품사 천이 확률을 추출하여 실제 자료와 비교하는 방식이라고 정리할 수 있다.

그러나, 이러한 종래의 확률기반적 구문 분석방법은 다음과 같은 한계를 가 진다. 첫째, 대량의 코퍼스가 인간이 만들어낼 수 있는 모든 다양한 구문 구조를 망라한다는 보장이 없기 때문에 이러한 한계를 부분적으로 극복하고자 특정 영역에 국한된 코퍼스를 구축할 수밖에 없다. 따라서 지식의 완결성이 보장되지 않으며, 사용 영역이 제한적일 수밖에 없다.

둘째, 오분석 자료가 발견되었을 때에 이에 대한 대처가 원천적으로 불가능하다. 즉, 확률을 사람의 손으로 수정할 수 없기 때문이다. 이를 수정하기 위해서는 새로운 코퍼스를 구축해야 하는 데, 일정 규모를 넘어설 경우, 확률은 더 이상 변동하지 않는 특성을 보인다.

특히, 이러한 종래의 확률기반적인 구문 분석방법을 한국어에 적용한 한국어 문법 모델은 크게 최현배(1937)에 기반한 전통적 모델과, 촘스키(Chomsky, 1965) 등에 기반한 생성 문법적 모델로 대별된다.

그러나, 이 두 가지의 모델에서는 구문 분석의 가장 기초적으로 요구되는 사항인 통사 단위를 확정하는 것이 비일관적이여서 만족스럽지 못하다. 즉, 전자는 조사를 단어로 취급하는 반면에 어미를 형태론적 단위로 처리하고, 이와는 반대로 후자는 조사(조사의 일부)를 형태론적 단위로 취급하는 반면 어미를 통사단위, 즉 단어로 취급하는 것이다.

따라서, 종래에는 주어진 입력 데이터를 구성하는 단위 표현들 사이의 의존 관계를 분석하여 이들의 문법적 기능을 포착하기 위하여 문법적 기능이 형상적 위치에 따라 결정된다는 양분지 구문 구조 방법이 사용되었다.

이러한, 양분지 구조(binary structure)를 설명하면, 만약 '나는 공원에서 영희를 만났다.(S)'라는 구문을 분석할 때, 문장을 구성하는 모든 단위가 둘씩 짝 지워져서 문장을 이루는 것으로서, '나는(NP), 공원에서 영희를 만났다.(VP)'로

나누고, VP를 다시, '공원에서(PP), 영희를 만났다.(V')'로 나누며, V'를 다시, '영희를(NP), 만났다.(V)로 나누는 지배관계와 선행관계가 하나의 규칙 속에서 동시에 정의되는 방식이다. 즉, 주어는 S에 직접 지배를 받는 NP이고, 처소는 VP에 직접 지배를 받는 PP이며, 직접 목적어는 V'에 직접 지배를 받는 NP인 것과 같이 이차적으로 문법적 기능이 정의된다.

이러한 종래의 양분지 구조에서는 문장의 직접 구성 성분들의 문법적 기능에 해당 성분이 구조 속에서 차지하는 위치에 따라 결정되고, 술어가 문장의 맨 뒤에 위치해야 한다는 한국어의 어순 제약을 준수하더라도 수학적으로 4개의 직접 구성 성분으로 이루어진 문장을 2개씩 묶어서 구조화한다면 수학적 가능성은 7가지($3 \times 2 \times 1 + 1$)가 되고, 5개의 성분으로 이루어진 문장은 무려 30가지($4 \times 3 \times 2 \times 1 + 2 \times 2$)의 중의적 구조를 만들어내서 구조적 중의성이 기하급수적으로 증가하게 된다.

즉, 한국어와 같은 자유 어순 언어의 경우는 물론이고, 고정어순 언어로 알려진 영어의 경우에도 전치사구는 어순 도치가 매우 자유롭기 때문에 이러한 어순도치로 인해 형상적 위치에 따라 문법적 기능이 결정될 수 없음을 보여준다.

또한, 기존의 양분지 구조로 분석할 경우, N개의 단위 표현으로 구성된 문장은, 2의 (n-2)승의 구조적 중의성이 발생한다. 즉 문장을 구성하는 어절수가 증가함에 따라 문장의 중의성이 기하급수적으로 증가한다.

또한 양분지 구조의 문제점은 성분의 자리바꿈이 일어나는 경우에 이를 예측할 수 있는 방법이 없다는 것이다. 한국어의 경우 직접 구성 성분의 수가 n일 때에 자리 바꾸기 가능성은 n!이 된다.

특히, 이러한 자유어순에 대한 대처능력은 문어 자료와는 달리 성분의 빈번한 생략과 자리바꿈이 있는 구어체 자료의 처리에 있어서 매우 중요한 것이나 종래의 양분지 구조는 이러한 자료까지 완벽하게 처리할 수 없었다.

따라서, 이러한 굴절어인 인구어(Indo-European language)를 기술하기 위한 종래의 구문 분석 모델은 교착어로서 판이한 언어유형을 보이는 한국어에 적합하지 못하고, 이러한 종래의 구문 분석 방법의 성공률은 태생적 한계에 의해 대략 50퍼센트 내지 60퍼센트 정도에 지나지 않는다.

특히, 이러한 종래의 구문 분석방법은, 구성 성분의 활용되는 형태에 따라 문법적 기능을 정의하는 활용 개념에 따른 것으로, 이러한 활용 개념에 따르면,

1a. 영희는 학교에 <u>간다</u> .

1b. 철수는 학교에 <u>가는</u> 영희를 보았다.

에서, (1a)의 '간다'나 (1b)의 '가는'은 모두 동사 '가다'의 활용형이다. 그런데, (1a)의 '간다'가 문장을 완성하는 반면에, (1b)의 '가는'은 문장을 종결짓는 것이 아니라 뒤에 오는 '영희'를 수식/한정한다. 따라서 종래의 문법에서는 '가는'과 같은 활용형을 '관형사형'이라고 부른다.

그러나, 이러한 종래의 입장에서 한 어휘가 동사이면서 동시에 관형사형이기도 하다면, 필연적으로 범주적 미결정성(categorial indeterminary)의 문제가 야기 된다. 즉, 문제의 '가는'이 '영희'를 수식하는 관형사라면 관형사는 '학교에'와 같은 성분을 이끌 수 없고, 만일 '가는'이 동사라면 문장을 완성하지 못하고 뒤에 오는 명사를 수식하는 지를 설명할 수 없는 것이다.

결국, 이를 설명하기 위해서는 '가는'이라는 활용형의 내부를 분석하여 어간 '가-'와 어미 '-는'의 구조를 참조할 수밖에 없으나 종래의 통사 규칙은 어휘 내부, 즉 활용형의 내부를 참조할 수 없기 때문에 엔진과 언어 지식사이에 독립성이 확보되지 못한다.

따라서, 이러한 종래의 구분 분석방법의 문제점들로 인하여 현재 상용화된 한국어 구문 분석방법이 없고, 실험실 수준의 구문 분석 방법만 시도되고 있을 뿐, 기계 번역의 경우에도 외-한 기계번역기가 주류를 이루고 있을 정도로 한국어 구문 분석에 대한 기술은 전무한 실정이다.

아울러, 종래의 구문 분석을 통한 기존의 자연어 검색 엔진들은 낮은 수준의 형태소 분석만을 이용하거나 어절 단위의 색인 방식을 사용함으로써 각각의 어절이 담고 있는 문법적 관계를 포착할 수 없고, 단지 확률기반적 접근에 따라 검색이 이루어져서 빈도 수만 높은 쓰레기 정보가 다량으로 검출되어 핵심적인 결과를 검색하기 어려웠다.

## 발명이 이루고자 하는 기술적 과제

상기와 같은 문제점을 해결하기 위한 본 발명의 목적은, 가속화되는 정보화 시대의 요구에 능동적으로 대응할 수 있는 다양하고 유용한 도구 개발에 필요한 핵심 기초 기술을 제공할 수 있고, 엄밀한 언어학적 성과에 기반함으로써 모든 영역 에 두루 사용할 수 있도록 강인성과 보편성 및 고신뢰도를 갖게 하며 언어 지식과 분석 엔진 사이의 독립성을 제고함으로써 지속적이고도 신속한 성능 개선이 가능하여 경제적인 측면에서도 매우 효율적으로 활용될 수 있는 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법을 제공함에 있다.

또한, 본 발명의 다른 목적은, 어떠한 어순 도치(scrambled)형 구문도 분석이 용이하고 별도의 어려운 분석 장치 없이 빠르게 처리할 수 있고, 어미를 어휘로 처리하여 구절구조 규칙(phrase structure rule)으로 이들의 결합을 제어함으로써 언어 모델과 분석 엔진 사이의 독립성을 제고할 수 있으며 각각에 대한 효율적인 개선을 가능하게 하는 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법을 제공함에 있다.

또한, 본 발명의 또 다른 목적은, 모빌형 구문분석기를 이용하여 성분 정보를 색인함으로써 문장을 구성하는 표현들 사이의 문법적 관계를 정확하게 포착해 낼 수 있으며, 그 결과 사람이 판단하는 바와 같은 방식으로 사용자가 요구하는 정보를 검색해서 정확한 정보를 제공할 수 있게 하는 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법을 제공함에 있다.

## 발명의 구성 및 작용

상기 목적을 달성하기 위한 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법은, 구문을 분석하여 구문의 문법적 기능을 명시하는 구문 분석방법에 있어서, 입력된 문장의 형태소를 분석하는 형태소 사전 프로그램, 문법 규칙이 저장되는 문법 규칙 데이터베이스, 조사와 어미를 통사의 단위로 취급하는 표지이론에 입각하여 용언 어미의 통사적 지위를 인정하여 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어간 및 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 구축하고, (a) 분석할 문장이 입력되면, 상기 형태소 사전 프로그램에 의해 어절별 형태소 내역을 분석하고, 어절별 형태소 분석자료 중 해당 입력자료에 적합한 형태소(품사)의 분석 사례를 선택하여 전처리를 실시하는 형태소 분석단계; 및 (b) 분석된 형태소들을 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 상기 하위범주화 데이터베이스를 이용하여 전체적인 구조를 확립하며, 각 구조의 가중치를 계산하여 가장 적합한 최적례를 확정하여 출력하는 구문 분석단계;를 포함하여 이루어지는 것을 특징으로 한다.

또한, 상기 구문 분석단계는, 다중 형태소 목록 프로그램에 의해 다중 형태소 목록에 포함된 구문이 존재하는 지를 판단하여 다중 형태소 구문이 존재하면 다중 형태소 형태로 변환하는 단계와, 의미자질 사전 프로그램에 의해 단어가 의미하는 의미를 판단하여 형태소에 포함시키는 단계를 구비하여 이루어지는 전처리 단계; 의미자질 품사가 부착된 형태소가 입력되면, 개별 형태소로 처리하고, 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 선택된 형태소에 국부 구조 규칙이 적용되는 지를 판단하여 국부적인 구조를 형성하고, 후속 처리대상을 참조하며, 재귀적 국부 구조가 형성되었는 지를 판단하여 내부 구조를 확립하는 내부 루프 가동단계 및 다른 내부 구조가 없으면 다음 프로세스를 반복하는 내부 루프 반복단계를 구비하여 이루어지는 부분구조 형성단계; 상기 하위범주화 데이터베이스와 첨어유형 데이터베이스를 기준으로 구문의 카테고리와 수식형태에 따라 전체 구조를 확립하는 전체구조 형성단계; 구문의 위치나 구문의 성격을 기준으로 각 구조의 가중치를 계산하고, 가장 중요한 구조를 선택하여 최적례를 선택하는 최적례 선택단계; 및 확정된 최적례의 전체 구조와 각각의 부분구조 및 각 형태소들간의 관계가 서로 짝을 이루어 연결되도록 모빌형(트리형) 연결선으로 표시하는 최적례 출력단계;를 포함하여 이루어지는 것이 바람직하다.

또한, 상기 의미자질사전 프로그램은, 논항의 핵어의 의미 정보를 확정하는 요소로서, 복문구조에서 구조적 중의성을 줄이는 데 기여하고, 용언별 첨어의 목록을 확정하도록 일반 명사 등 단어가 의미하는 의미와 그것들에 대한 분류를 유형별로 실시하는 프로그램이고, 상기 다중 형태소목록 프로그램은, 서로 동일한 형태의 조사나 조사의 기능을 갖는 접미사 등에 대한 어휘적인 특징을 분류하기 위해서 구별을 위한 분류를 유형별로 실시하는 프로그램이며, 상기 문법 규칙 데이터베이스는, 각 기본소에 대한 문법적인 규칙을 규정하는 정보를 저장하는 것이고, 상기 하위범주화 데이터베이스는, 용언이 취할 수 있는 성분의 내역 및 변형가능한 용언 어미의 형태에 대한 정보를 저장하는 것이며, 상기 첨어유형 데이터베이스는, 다분지 구조의 중의성을 결정하는 요소로서, 조사나 조사의 기능을 갖는 접미사 등에 대한 일반적인 특징에 대한 정보를 저장하는 것이 바람직하다.

한편, 상기 목적을 달성하기 위한 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법은, 자연어 질의어를 입력하여 문서(문장)를 검색하는 자연어 검색 방법에 있어서, 용언 어미의 통사적 지위를 인정하여 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 구축하고, 분석할 문장이 입력되면, 형태소를 분석하고, 분석된 형태소들을 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 상기 하위범주화 데이터베이스를 이용하여 전체적인 구조를 확립하는 모빌적 형상 개념을 기초로 한 구문 분석방법으로 검색 대상이 되는 문서의 문장분석 정보를 문장정보 데이터베이스에 저장하는 문서 분석단계; 상기 문서정

보 데이터베이스에서 원하는 정보를 질문하는 자연어 형태의 질의어가 입력되면, 상기 모빌적 형상 개념을 기초로 한 구문 분석방법으로 질의어의 구문을 먼저 분석하고, 분석된 구문 분석 결과를 구문 정보에 의해 단어별로 해부하며, 질의어의 의문문 형태를 파악하여 해부된 세부 질의어를 확정하는 질의어 구문 분석단계; 상기 문장분석 사전에서 확정된 상기 세부 질의어의 태그를 원하는 의문문의 형태에 따라 검색용 태그로 역할 변환하고, 변환된 검색용 태그를 갖는 단어를 상기 문장분석 사전에서 검색하여 검색된 횟수를 기준으로 순위를 계산하는 문서 검색단계; 및 검색된 단어와 및 검색용 태그를 포함하는 문장과 그 문장이 포함된 문서에 대한 내용을 표시하는 결과 표시단계;를 포함하여 이루어지는 것을 특징으로 한다.

이하, 본 발명의 바람직한 일 실시예에 따른 모빌적 형상 개념을 기초로 한 구문 분석방법 및 이를 이용한 자연어 검색 방법을 도면을 참조하여 상세히 설명한다.

먼저, 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법은, 표지이론 에 입각하여 용언 어미의 통사적 지위를 인정하여 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어간 및 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 기준으로 구문을 분석하는 방법이다. 즉, 이러한 구문분석 방법은 고유한 한국어 문법 모델과 언어학적 지식을 컴퓨터에 직접 입력한 것으로, 모든 언어에 적용될 수 있다는 점에서 지식기반적(knowledge-based approach)이라 할 수 있다. 상기 하위범주화 데이터베이스의 일례는 이하 각 단계별 설명에서 후술될 것이다.

이러한 표지이론의 핵심 문법 모델은, 조사와 어미를 모두 통사의 단위, 즉 하나의 단어로 취급하는 것으로서, 예를 들어, 상술된 활용 개념에서 '영희는 학교에 간다.'와, '철수는 학교에 가는 영희를 보았다'라는 문장이 있을 때, 표지이론은,

2a. [영희 - 는 학교 - 에 가] - ㄴ - 다.

2b. [철수 - 는 [학교 - 에 가] - 는 영희 -를 보] - 았 - 다.

와 같이 '가는'의 '-는'이나 '간다'의 '-ㄴ-', '-다'가 모두 표지이며 통사적 단위로 구분된다. 그리고, 각각의 표지가 담당하는 기능은 서로 다르다.

즉, '가는'의 '-는'은 동사구를 명사와 통합시키는 역할을 하지만, '간다'의 '-ㄴ-'은 현재(진행)상을, 그리고 '-다'는 서술의 서법을 나타낸다. 이렇게 함으로써 어휘들 사이의 통합 관계가 온전히 문법에서 규정될 수 있고, 이에 따라 문법과 분석 엔진 사이의 독립성이 제고됨으로써 오분석 자료의 발견이나 수정도 용이해진다.

또한, 지배 관계(dominance relation)와 선후 관계(precedence relation)를 구분하는 ID-LP format을 채택한 모빌적 형상(mobile configuration)을 채택함으로써 동일한 성분으로 이루어졌으나 성분들의 순서만 바뀐 문장들을 동일하게 분석해 낼 수 있는 것이다.

이러한 표지이론에 입각한 본 발명의 본 발명의 바람직한 일 실시예에 따른 모빌적 형상 개념을 기초로 한 구문 분석방법은, 도 1에 도시된 바와 같이, 구문을 분석하여 구문의 문법적 기능을 명시하기 위한 구문 분석방법으로서, 어순이 도치된 문장의 분석이 가능하도록 조사 및 어미를 독립된 단어로 판단하여 형태소의 문법적 기능과 특징을 데이터베이스에 미리 저장하고, 분석이 필요한 구문이 입력되면, 각 성분의 중심어(Head)가 가지는 엄밀한 하위범주화 내역을 기반으로 여기에 포함된 의미 자질(semantic feature) 및 조사 형태, 그리고 범주 정보(categorial identity)를 근거로 구문분석을 시도함으로써 과생성을 억제하고, 하위범주화 정보에 미리 규정되어 있는 문법적 기능(grammatical role) 정보를 기준으로 각 형태소들간의 관계를 특정 기호로 명기하여 구문의 문법적 관계를 명시하는 것으로서, 크게 형태소 분석단계(S1)(S2)(S3) 및 구문 분석단계(S4)(S5)(S6)(S7)(S8)(S9)(S10)로 이루어지는 구성이다.

즉, 본 발명의 형태소 분석단계는, 우선, 조사나 용언 어미를 독립된 기본소로 판단하여 어미의 문법적 기능의 특징이 형태소 사전의 형태로 저장되는 형태소 사전 프로그램(1), 문법적 규칙이 저장되는 문법 규칙 데이터베이스(4)를 구축하고, 분석할 문장이 입력되면(S1), 상기 형태소 사전 프로그램(4)에 의해 구문의 최 소단위인 형태소를 분석하고(S2), 여기에 품사에 태그를 달아 구분하는 품사 부착단계(S3)들로 이루어진다.

여기서, 분류된 형태소들은 문법적 기능을 표시하는 태그 및 약자가 첨부되는 것으로서, 도 4의 구문 분석결과창의 오른쪽 창에 도시된 바와 같이, 주어와 주격 조사, 목적어와 목적격 조사, 서술어와 서술어미 등의 형태로 의미를 갖는 최소단위인 형태소로 분류하고, 각 형태소에 태그를 달아 형태소의 종류를 약자(np, jc, pv 등)를 기재하여 표시한다.

이어서, 본 발명의 구문 분석단계(S4)(S5)(S6)(S7)(S8)(S9)(S10)는, 구분된 형태소들을 문법 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 수식의 형태에 따라 전체적인 구조를 확립하며, 각 구조의 가중치를 계산하여 최적례를 확정하여 각 형태소들간의 관계를 특정 기호로 명기하고 구문의 문법적 관계를 명시하는 것으로서, 도 1에 도시된 바

와 같이, 전처리단계(S4)와, 부분 구조 형성단계(S5)와, 전체 구조 형성단계(S6)(S7) 및 전체구조 확정단계(S7)(S8)(S9)(S10)로 이루어지는 구성이다.

여기서, 상기 전처리단계(S4)는, 도 2에 도시된 바와 같이, 품사가 태그된 형태소가 입력되면(S41)에 다중 형태소 목록 프로그램(3)에 의해 다중 형태소 타입의 구문이 존재하는 지를 판단하여(S42) 다중 형태소 구문이 존재하면 다중 형태소 형태로 변환하는 단계(S43)와, 의미자질 사전 프로그램(2)에 의해 형태소의 의미를 판단하여 의미 자질에 대한 형태소가 필요하면(S44) 의미 자질 형태소를 추가시키는 단계(S45)를 구비하여 이루어진다.

이때, 상기 의미자질사전 프로그램(2)은, 아래에 예시된 바와 같이, 논항의 핵어의 의미 정보를 확정하는 요소로서, 복문구조에서 구조적 중의성을 줄이는 데 기여하고, 용언별 첨어의 목록을 확정하도록 일반 명사 등 단어가 의미하는 의미와 그것들에 대한 분류를 유형별로 실시하는 것이다.

<의미자질사전 프로그램의 적용례>

@root 밥

@pos nc

@type concrete

@subtype food

@property solid

.....

@root 학교

@pos nc

@type concrete|abstract

@subtype organization

.....

또한, 상기 다중 형태소 목록 프로그램(3)은, 아래 예시된 바와 같이, 서로 동일한 형태의 조사나 조사의 기능을 갖는 접미사 등에 대한 어휘적인 특징을 분류하기 위해서 구별을 위한 분류를 유형별로 실시하는 것이다.

<다중 형태소 목록 프로그램 적용례>

jc <- 에/jc 대/nx-하/xsv-어서/ec

.....

jc <- 와/jc 같/pa-이/xsa

.....

pv <- */nc-*/xsv

pv <- */nx-*/xsv

nc <- */nc-*/nx

.....

ep <- ??/etm-것/nb-이/co

{ep:tense=[fut]; ep:origin = [cep];}

.....

이어서, 상기 부분 구조 형성단계(S5)는, 도 3에 도시된 바와 같이, 상기 의미자질 품사 부착 형태소가 입력되면(S51) 개별 형태소들을 처리하고(S52), 문법 규칙 데이터베이스(4)에 저장된 문법적 규칙에 따라 국부 구조가 존재하는 지를 판단하여(S53) 국부 구조를 형성하고(S54), 후속 처리 대상을 참조하여(S55) 재귀적 국부 구조를 형성한다.(S56) 이러한 재귀적 국부 구조는 다시 부분적인 국부 구조를 확립하여 국부 구조를 확립하는 내부 루프 가동단계(S53)(S54)(S55)(S56) 및 다른 국부 구조가 없으면 다음 형태소를 선택하여 반복하는 내부 루프 반복단계(S57)를 구비하여 이루어진다.

여기서, 상기 문법 규칙 데이터베이스(4)는, 아래 예시된 바와 같이, 각 기 본소에 대한 문법적인 규칙을 규정하는 정보를 저장하는 것이다.

<규칙사전 용례>

N' <- NPm N' <5>

[NPm:nbval;]

{N':type = N'#1:type;

N':subtype = N'#1:subtype;

N':property = N'#1:property;}

.....

ADVP <- mag ADVP-s <4>

[s:lex == [,]; mag:subtype ** [degree];]

{ADVP:subtype = ADVP#1:subtype;}

.....

이어서, 도 1에 도시된 바와 같이, 상기 전체 구조 형성단계(S6)(S7)는, 하위 범주화 데이터베이스(5)와 첨어 유형 데이터베이스(6)를 기준으로 구문의 카테고리와 수식형태에 따라 전체적인 구조를 형성하는 단계(S6)와, 또 다른 형태의 유효 매트릭스의 검사 여부를 판단하여(S7) 다음 매트릭스의 부분 구조 형성단계(S5)를 반복하는 단계들로 이루어진다.

여기서, 상기 하위 범주화 데이터베이스(5)는, 조사와 어미를 모두 통사의 단위로 취급하는 표지이론에 입각하여 용언 어미의 통사적 지위를 인정하고, 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어간 및 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 것으로서, 아래에 예시한 바와 같이, 예를 들어 중심어 '먹다'에서 '먹-'의 변형 가능한 용언 어미의 형태에 대한 정보를 저장하는 것이다.

<하위범주화 데이터베이스 적용례>

먹 NP(subtype ~= [human|animal]; jcval *= <이>)[c_sbj]

NP(type ~= [concrete]; subtype ~= [food|medicine|abstract|fuel]; jcval *= <을>)[c_obj]

{A_Type1}

pv

.....

먹이 NP(jcval *= <이>; !!(nbval); type ~= [alive])[c_sbj]

NP(jcval *= <에게>; type ~= [alive])[c_dat]

NP(jcval *= <을>; subtype ~= [food|liquid])[c_obj]

{A_Type1}

pv

.....

또한, 상기 첨어 유형 데이터베이스(6)는, 다분지 구조의 중의성을 결정하는 요소로서, 조사나 조사의 기능을 갖는 접미사 등에 대한 일반적인 특징에 대한 정보를 아래에 예시된 바와 같이 저장하는 것이다.

<첨어 유형 데이터베이스 적용례>

#BOAT

A_Type1

ADVP(subtype ** [manner])[a_manner]

ADVP(subtype ** [time])[a_temp]

ADVP(subtype ** [motive])[a_reason]

...

NP(subtype ** [time]; !!(jcval) amp;amp nbval)[a_occurrence]

NP(subtype ~=[place|space|spot]; jcval**<에서>)[a_loc]

NP(type ** [concrete]; jcval**<로>)[a_instr]

...

VPn(etnval == [기]; jcval == [에])[a_motive]

VPf(mood ~= [declarative]; jcval == [고])[a_reason]

A_Type2

.....

A_Type3

.....

.....

#BOAT

이어서, 도 1에 도시된 바와 같이, 전체 구조 확정단계(S7)(S8)(S9)(S10)는, 구문의 위치나 구문의 성격을 기준으로 각 구조의 중요도를 기준으로 가충치를 계 산하고(S7), 가장 최적의 최적례를 선택하여(S8) 선택된 최적례를 출력하

는 단계(S10)를 구비하여 이루어진다.

이러한 최적례 출력 단계(S10)는, 도 4의 구문 분석결과창의 왼쪽 창에 도시된 바와 같이, 확정된 전체 구조와 각각의 내부구조와 외부구조 및 각 형태소들간의 관계가 서로 짝을 이루어 연결되도록 모빌형(트리형) 연결선으로 표시하는 단계이다.

그러므로, 이러한 한국어에 맞게 개발된 문법 모델과 언어학적 지식에 의거함으로써 종래의 확률기반적인 방식에 비하여 현격한 정확도를 보장하고, 사람의 언어 인식방식과 동일하기 때문에 단문 차원에서는 지식구축의 정도에 따라 원리상 100%에 가까운 처리율을 기대할 수 있다.

또한, 모빌적 형상을 채택함으로써 어순이 뒤바뀐 문장도 정확하고 일관되게 분석할 수 있고, 모든 언어 영역에 적용이 가능하여 영역(domain) 변경에 따른 추가적 비용이 발생하지 않으며, 다분지 구조를 채택하여 불필요한 분석을 줄이고, 이에 따라 오류 발생 원인의 파악이 용이하며, 지식과 엔진 사이의 독립성이 높으므로 오분석 자료에 대한 개선이 신속하게 이루어질 수 있다.

또한, 기하급수적으로 증가하는 종래의 양분지 구조의 중의성과는 달리 문법적 기능을 기본소로 하는 다분지 구조의 분석으로 말미암아 구조적 중의성이 단지 어절수의 증가에 따라 산술급수적으로 증가하여 구문 분석이 용이하고, 빈번한 생략과 자리바꿈이 일어나는 구어자료를 완벽히 분석할 수 있다.

한편, 이러한 모빌적 형상 개념을 기초로 한 구문 분석방법을 구현할 수 있 는 구문 분석기는, 각종 입출력장치를 제어하는 마이크로 프로세서나 CPU 등의 제어부와, 램(RAM)이나 롬(ROM)이나 하드디스크 등 각종 정보를 저장하는 저장장치를 구비하여 이루어지는 것으로서, 상기 제어부는, 도 1의 상기 형태소 사전 프로그램(1)과, 의미 자질 사전 프로그램(2)과 다중 형태소 목록 프로그램(3)을 포함하고, 상기 저장장치는, 문법적 규칙이 저장되는 문법 규칙 데이터베이스(4) 및 상기 하위 범주화 데이터베이스(5)와, 상기 첨어 유형 데이터베이스(6)를 포함한다.

즉, 상기 제어부는, 분석할 문장이 입력되면, 상기 형태소 사전 프로그램(1)에 의해 구문의 최소단위인 형태소를 분석하고, 구분된 형태소들을 상기 문법 규칙 데이터베이스(4)에 저장된 문법적 규칙에 따라 문장의 부분 구조를 먼저 확립하며, 상기 하위범주화 데이터베이스(5)에 저장된 하위 범주화 정보를 기준으로 전체적인 구조를 확립하여 각 구조의 가중치를 계산하고, 최적례를 선택하여 각 형태소들간의 관계를 특정 기호로 명기하고 그 구문의 문법적 관계를 명시하도록 프로그램된다.

따라서, 본원 발명의 구문 분석기는 그 문법적 기능(grammatical role)을 형상(configuration)에서 유추해 내는 방식이 아니라, 문법적 기능 자체를 기본소(primitives)로 보고, 미리 입력된 하위범주화 정보(subcategorization)를 이용하여 문법적 기능을 명시하는 방식을 채택한 것이다.

또한, 이러한 하위범주화 정보는 단순히 품사 목록만을 제공해서는 부족한 것으로서, 본 발명의 구문 분석기는 각각의 성분에 의미 정보를 기술함으로써 중의 성을 제거하고 최소한의 문법적인 구조만이 생성되도록 한다. 이를 위해서 상기 형태소 분석 단계(S1)(S2)(S3)에서 각각의 어휘가 가지는 의미 자질이 제시되도록 시스템을 설계하고 이를 통해 가능한 문법적 관계를 정확하게 파악해 낼 수 있다.

또한, 각각의 하위범주화 틀(subcategorization frame)들은 저마다 허용가능한 수식어 유형(adjunct type)을 요구한다. 따라서 이를 수식형태에 따라 전체적인 외부 구조를 확립하는 단계(S6)에서 기술해 줌으로써 불필요한 중의적 구조가 생성되는 것을 차단하고 적절한 구문분석이 이루어지게 한다.

한편, 이러한 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법은, 자연어 형태의 질의어를 입력하여 문서 또는 문장을 검색하여 원하는 지식을 찾아주는 검색 방법으로서, 도 5에 도시된 바와 같이, 크게 도 1에 도시된 바와 같이, 상기 구문 분석방법을 이용하여 입력된 문서를 분석하는 문서 분석단계(S1)~(S10)과, 질의어 구문 분석단계(S100)(S110)(S120)와, 문서 검색 단계(S130)(S140)(S150)(S160)(S170)(S180) 및 결과 표시단계(S190)(S200)(S210)(S220)를 포함하여 이루어지는 구성이다.

즉, 상기 문서 분석단계는, 도 1에 도시된 바와 같이, 문장을 입력하는 것이 아니라 문서를 입력하는 것으로서, 형태소의 문법적 기능과 특징을 데이터베이스에 미리 저장하고, 분석이 필요한 구문이 입력되면, 기본소를 이용하여 형태소들을 정의하고, 정의된 형태소에서 어미로 정의된 형태소와 일치하는 상기 데이터베이스의 문법적 지배관계에 따라 각 형태소들간의 관계를 특정 기호로 명기하여 구문의 문법적 관계를 명시하는 모빌적 형상 개념을 기초로 한 구문 분석방법으로 검색 대상 이 되는 문서의 문장분석 정보를 문장분석 사전(Dictionary)의 형태로 색인 데이터베이스에 저장하는 단계로서, 이는 상술된 구문 분석방법과 동일하다.

이러한 준비단계를 마치고, 도 5에 도시된 바와 같이, 상기 질의어 구문 분석단계(S110)(S120)는, 원하는 정보를 질

문하는 자연어 형태의 질의어가 입력되면(S100), 상술된 모빌적 형상 개념을 기초로 한 구문 분석방법으로 질의문의 구문을 분석하고(S110), 분석된 구문 분석 결과를 분석하여 구문 정보에 의해 단어별로 해부하며, 질의문의 의문문 형태를 파악하여 미리 입력된 문장정보 데이터베이스(10)의 세부 질의어를 기준으로 질의어를 확정하는 단계(S120)이다.

여기서, 자연어 형태의 질의문이란, 인간의 사고방식을 기준으로 인간이 쉽게 알아들을 수 있는 인간의 언어로서, 도 6의 상단 '검색어' 창에 예시된 바와 같이, 예를 들어 '누가 철수를 좋아하니?'와 같은 문장이다.

따라서, 이러한 질의어 구문 분석단계를 거쳐서 도 6의 질의어 분석결과(Query Analyzer), '누가 철수를 좋아하니?'의 구문을 'SUB(주어) OBJ(목적어) HEAD(서술어)'로 정의할 수 있다.

참고로, 도 6의 중단 '전체 색인량'창에는 상기 문서분석 단계에서 미리 분석된 문서의 개수가 '47건', 문장의 개수가 '92건', 단어의 개수가 '257건'임을 나타낸다.

이어서, 상기 문서 검색단계 중 문장 유형 판별 단계(130)는, 사전 데이터베이스(13)를 대상으로 상기 사전에서 확정된 상기 세부 질의어의 태그를 원하는 의문문의 형태에 따라 검색용 태그로 역할 변환하고, 변환된 검색용 태그를 갖는 단 어를 상기 사전 데이터베이스(13)에서 검색하는 단계이다.(S130)

즉, 도 6에 도시된 바와 같이, 의문문의 형태를 분석(WH-Analyzer)하여 '누가 => 의문사, 주어'로 도출하고, 이에 따라 검색용 태그의 역할이, 목적어였던 '철수를'을 그대로 목적어 또는 주어로 변환하여 태그를 '철수/nc'로 변환하고, 의문 서술어였던 '좋아하니?'를 일반 서술어로 변환하여 '좋아하/pv'로 변환하여 문장분석 사전(Dictionary)에서 검색한다.

여기서, 상기 문서 검색단계(130)는 검색자의 선택에 따라(140) 특별검색 규칙정보(11)와 명사체계 데이터베이스(12)에 의해 특별 검색 모드를 위한 조건을 발생시키는 특별 검색 모드 조건 생성단계(S150)를 거치거나, 사전 데이터베이스(13)를 일반 검색하는 일반 검색모드 조건 생성 단계(S160)를 거칠 수 있다.

일반 검색 모드란, 구문 분석된 정보만을 이용하여 질의어의 구문 분석 결과만을 기반으로 기 분석된 문서 데이터베이스를 검색하여 일치하는 내용을 추출, 제공하는 검색 방식을 말한다.

이러한 일반 검색 모드는, 주어진 질의어의 직접 구성 성분과 일치하는 자료를 추출하여 제공하는 성분 일치 검색 방법과, 질의어를 구성하는 성분들을 포함하되 핵어인 술어와 의미적으로 유사한 술어들을 포함하는 자료를 추출하여 제공하는 의미 일치 검색 방법을 사용할 수 있다.

또한, 특별 검색 모드란, 질의어에 특정한 표현이 포함되는 경우, 이를 기반으로 의미적으로 주어진 성분에 종속된 내용들을 검색하여 제공하는 방식으로서, 예컨대, '철수가 무슨 과일을 먹었니?'라는 질의어가 들어오면 찾고자 하는 문장 은'철수가 사과를 먹었다.' 등을 포함하여 철수가 특정 종류의 과일을 먹었다는 내용을 포함하는 문서를 추출하여 제공하는 검색 방식이다.

즉, 이러한, 특별 검색 모드를 위해서는 상기 특별검색 규칙정보(11)와 명사체계 데이터베이스(12)와 같은 명사의 의미적 위계 구조에 대한 데이터베이스가 사용된다.

이어서, 도 8에 도시된 바와 같이, 역할이 반전된 역파일 데이터베이스(14)의 데이터를 생성하기 위하여 접근하여 결과를 반환하고(S170), 다중 결과의 AND나 OR 조건으로 변환된 검색용 태그를 갖는 단어가 검색된 횟수를 도 9에 도시된 바와 같이, 연산한다(S180).

즉, 도 9 및 도 10에 도시된 바와 같이, 1번 문서에서 1번째 문장 '영희는 철수를 좋아한다.', 23번째 문장 '영희는 철수를 좋아한다.', 60번째 문장 '영희는 철수를 좋아한다.'가 검색되었다.

이어서, 상기 결과 표시단계(S190)(S200)(S210)(S220)는, 도 11에 도시된 바와 같이, 검색된 단어와 및 검색용 태그를 포함하는 문장과 그 문장이 포함된 문서에 대한 정보 및 내용 등 복수의 결과를 판별하여(S190) 빈도에 따라 순위를 계산하고(S200), 이를 포함한 문서 정보 데이터베이스(15)를 읽어서 외부 정보를 참조한 후,(S210) 이러한 결과를 출력하는 단계이다(S220).

따라서, 도 12에 도시된 바와 같이, 검색어 창에 '누가 철수를 좋아하니?'와 같이 자연어를 질의어로 입력하면, 질의어 구문 분석결과 창에 조사와 어미를 형태소로 분석하여 '누/np', '가/jc', '철수/nc', '를/jc', '좋아하/pv', '니/et', '?/s'와 같이 표시하고, 이를 검색용 태그를 갖는 단어로 검색하여 그 결과를 검색 결과창에 나타나고, 이러한 검색 결과창에는 '영희는 철수를 좋아한다.'와 같은 문장과 함께 질문자의 복합적인 판단이 가능하도록 '그리고 철수는 순자도 좋아

한다.'와 같은 문장을 표시할 수 있다.

한편, 도시하진 않았지만, 이러한 자연어 검색 방법을 이용한 자연어 검색시스템은, 각종 입출력장치를 제어하는 마이크로 프로세서나 CPU 등의 제어부와, 램(RAM)이나 롬(ROM)이나 하드디스크 등 각종 정보를 저장하는 저장장치를 구비하여 이루어지는 것으로서, 상기 저장장치는, 형태소의 문법적 기능과 특징을 데이터베이스에 미리 저장하고, 분석이 필요한 구문이 입력되면, 기본소를 이용하여 형태소들을 정의하고, 정의된 형태소에서 어미로 정의된 형태소와 일치하는 상기 데이터베이스의 문법적 지배관계에 따라 각 형태소들간의 관계를 특정 기호로 명기하여 구문의 문법적 관계를 명시하는 모빌적 형상 개념을 기초로 한 구문 분석방법으로 검색 대상이 되는 문서의 문장분석 정보를 문장분석 사전(Dictionary)의 형태로 색인 데이터베이스가 구축되는 것이다.

또한, 상기 제어부는, 색인 데이터베이스에서 원하는 정보를 질문하는 자연어 형태의 질의문이 입력되면, 상기 모빌적 형상 개념을 기초로 한 구문 분석방법으로 질의문의 구문을 분석하고, 분석된 구문 분석 결과를 구문 정보에 의해 단어별로 해부하며, 질의문의 의문문 형태를 파악하여 해부된 문장분석 사전용 세부 질의어를 확정하고, 상기 문장분석 사전에서 확정된 상기 세부 질의어의 태그를 원하는 의문문의 형태에 따라 검색용 태그로 역할 변환하고, 변환된 검색용 태그를 갖 는 단어를 상기 문장분석 사전에서 검색하여 검색된 횟수를 카운팅하며, 검색된 단어와 및 검색용 태그를 포함하는 문장과 그 문장이 포함된 문서에 대한 내용을 빈도 순위에 따라 표시하도록 프로그램되는 것이다.

따라서, 본 발명에 의해 구현된 자연어 검색 시스템은 색인할 문서를 수집한 후 각각의 문서를 구성하는 문장들을 색인하고 각 문장은 다시 구문분석기가 출력하는 결과에 따라 구성 성분별로 문법적 기능을 색인함으로써 유관한 정보를 포함하는 문서가 있기만 하면 정확하게 그 정보가 들어있는 문서를 찾아 제시할 수 있다.

예를 들면, 도면에 예시된 '누가 철수를 좋아하니?' 이외에도, '철수가 누구를 만났니?' 혹은 '철수가 만난 사람은?' 과 같은 질의어가 입력되면 '만나다'에서 질문의 초점이 목적어에 있으므로 '만나다'라는 술어에 대해 주어가 '철수'이고 목적어가 있는 질의어로 문장을 검색하여 그 결과를 제공할 수 있다.

그러므로, 의미 정보를 포함하기 때문에 문장형 질의어의 경우 유사어를 자동으로 확정함으로써 신속하고 정확성이 높은 검색이 가능해지고, 의미 연산까지 포함하는 지능적 검색이 가능하다.

또한, 검색 결과에 대한 연관성을 현격하게 향상시킬 수 있고, 단순한 일치의 검색을 넘어서서 문법적 관계까지 고려한 정확하고 지능적인 검색이 가능하다.

또한, 이러한 구문 분석과 자연어 검색을 기반으로 한-외 기계번역기 시장을 새롭게 창출하고, 그밖에도 지능적 언어처리를 위한 다양한 분야의 시장이 새로 형성될 수 있는 것이다.

본 발명은 상술한 실시예에 한정되지 않으며, 본 발명의 사상을 해치지 않는 범위 내에서 당업자에 의한 변형이 가능함은 물론이다.

예컨대, 본 발명의 실시예에서는 한국어에만 국한되어 있으나 조사나 어미의 중요성이 높은 일본어 등 다른 나라의 언어에도 적용될 수 있고, 구문 분석기를 이용한 자연어 검색 시스템은 물론, 야후 등의 포탈 사이트의 검색 엔진이나, 인공지능 컴퓨터의 질문, 응답시스템 등 컴퓨터가 인간의 언어를 이해할 수 있는 모든 분야에 적용될 수 있는 것이다.

따라서, 본 발명에서 권리를 청구하는 범위는 상세한 설명의 범위 내로 정해지는 것이 아니라 후술되는 청구범위와 이의 기술적 사상에 의해 한정될 것이다.

## 발명의 효과

이상에서와 같이 본 발명의 모빌적 형상 개념을 기초로 한 구문 분석방법과, 이를 이용한 자연어 검색 방법에 의하면, 다양하고 유용한 인터페이스 도구 개발에 필요한 핵심 기초 기술을 제공할 수 있고, 모든 컴퓨터 영역에 두루 사용할 수 있도록 강인성과 보편성을 갖게 하며, 지속적이고도 신속한 성능 개선이 가능하여 경제적인 측면에서도 효율적이고, 어떠한 어순 도치형 구문도 분석이 용이하여 별도의 어려운 분석 장치 없이 빠르게 처리할 수 있으며, 문장을 구성하는 표현들 사이의 문법적 관계를 정확하게 포착해 낼 수 있으며, 그 결과 사람이 판단하는 바와 같은 방식으로 사용자가 요구하는 정보를 검색해서 정확한 정보를 제공할 수 있는 효과를 갖는 것이다.

## (57) 청구의 범위

### 청구항 1.

구문을 분석하여 구문의 문법적 기능을 명시하는 구문 분석방법에 있어서,

입력된 문장의 형태소를 분석하는 형태소 사전 프로그램, 문법 규칙이 저장되는 문법 규칙 데이터베이스, 조사와 어미를 모두 통사의 단위로 취급하는 표지이론에 입각하여 용언 어미의 통사적 지위를 인정하고, 어휘들 사이의 통합관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어간 및 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 구축하고,

(a) 분석할 문장이 입력되면, 상기 형태소 사전 프로그램에 의해 어절별 형태소 내역을 분석하고, 어절별 형태소 분석자료 중 해당 입력자료에 적합한 형태소를 선택하여 전처리를 실시하는 형태소 분석단계; 및

(b) 분석된 형태소들을 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 상기 하위범주화 데이터베이스를 이용하여 전체적인 구조를 확립하며, 각 구조의 가중치를 계산하여 가장 적합한 최적례를 확정하여 출력하는 구문 분석단계;

를 포함하여 이루어지는 것을 특징으로 하는 모빌적 형상 개념을 기초로 한 구문 분석방법.

**청구항 2.**
제 1항에 있어서,

상기 구문 분석단계는,

다중 형태소 목록 프로그램에 의해 다중 형태소 목록에 포함된 구문이 존재하는 지를 판단하여 다중 형태소 구문이 존재하면 다중 형태소 형태로 변환하는 단계와, 의미자질 사전 프로그램에 의해 단어가 의미하는 의미를 판단하여 형태소에 포함시키는 단계를 구비하여 이루어지는 전처리 단계;

의미자질 품사가 부착된 형태소가 입력되면, 개별 형태소로 처리하고, 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 선택된 형태소에 국부 구조 규칙이 적용되는 지를 판단하여 국부적인 구조를 형성하고, 후속 처리대상을 참조하며, 재귀적 국부 구조가 형성되었는 지를 판단하여 내부 구조를 확립하는 내부 루프 가동 단계 및 다른 내부 구조가 없으면 다음 프로세스를 반복하는 내부 루프 반복단계를 구비하여 이루어지는 부분구조 형성단계;

상기 하위범주화 데이터베이스와 첨어 유형 데이터베이스를 기준으로 구문의 카테고리와 수식형태에 따라 전체 구조를 확립하는 전체구조 형성단계;

구문의 위치나 구문의 성격을 기준으로 각 구조의 가중치를 계산하고, 가장 중요한 구조를 선택하여 최적례를 선택하는 최적례 선택단계; 및

확정된 최적례의 전체 구조와 각각의 부분구조 및 각 형태소들간의 관계가 서로 짝을 이루어 연결되도록 모빌형(트리형) 연결선으로 표시하는 최적례 출력단계;

를 포함하여 이루어지는 것을 특징으로 하는 모빌적 형상 개념을 기초로 한 구문 분석방법.

**청구항 3.**
제 2항에 있어서,

상기 의미자질사전 프로그램은, 형태소의 통사 특성과 의미 정보를 확정하는 요소로서, 복문구조에서 구조적 중의성을 줄이는 데 기여하고, 용언별 첨어의 목록을 확정하도록 일반 명사 등 단어가 의미하는 의미와 그것들에 대한 분류를 유형별로 실시하는 프로그램이고, 상기 다중 형태소목록 프로그램은, 서로 동일한 형태의 조사나 조사의 기능을 갖는 접미사 등에 대한 어휘적인 특징을 분류하기 위해서 구별을 위한 분류를 유형별로 실시하는 프로그램이며, 상기 문법규칙 데이터베이스는, 각 기본소에 대한 문법적인 규칙을 규정하는 정보를 저장하는 것이고, 상기 하위범주화 데이터베이스는, 용언이 취할 수 있는 성분의 내역 및 변형 가능한 용언 어미의 형태에 대한 정보를 저장하는 것이며, 상기 첨어 유형 데이터베이스는, 다분지 구조의 중의성을 결정하는 요소로서, 핵어에 의해 통합될 수 있는 국부구조의 유형을 결정하는 조사나 어미 또는 이들과 유사한 기능을 갖는 접미사 등에 대한 일반적인 특징에 대한 정보를 저장하는 것을 특징으로 하는 모빌적 형상 개념을 기초로 한 구문 분석방법.

**청구항 4.**
자연어 질의어를 입력하여 문서(문장)를 검색하는 자연어 검색 방법에 있어서,

용언 어미의 통사적 지위를 인정하여 어휘들 사이의 통합 관계가 온전히 문법적으로 규정될 수 있도록 문장 각 구성 성분의 어미 등 중심어(head)가 가지는 하위범주에 대한 내역이 저장되는 하위범주화 데이터베이스를 구축하고, 분석할 문 장이 입력되면, 형태소를 분석하고, 분석된 형태소들을 문법 규칙 데이터베이스에 저장된 문법적 규칙에 따라 문장의 부분적인 구조를 먼저 확립하고, 상기 하위범주화 데이터베이스를 이용하여 전체적인 구조를 확립하는 모빌적 형상 개념을 기초로 한 구문 분석방법으로 검색 대상이 되는 문서의 문장분석 정보를 문장정보 데이터베이스에 저장하는 문서 분석단계;

상기 문서정보 데이터베이스에서 원하는 정보를 질문하는 자연어 형태의 질의어가 입력되면, 상기 모빌적 형상 개념을 기초로 한 구문 분석방법으로 질의어의 구문을 먼저 분석하고, 분석된 구문 분석 결과를 구문 정보에 의해 단어별로 해부하며, 질의어의 의문문 형태를 파악하여 해부된 세부 질의어를 확정하는 질의어 구문 분석단계;

상기 문장분석 사전에서 확정된 상기 세부 질의어의 태그를 원하는 의문문의 형태에 따라 검색용 태그로 역할 변환하고, 변환된 검색용 태그를 갖는 단어를 상기 문장분석 사전에서 검색하여 검색된 횟수를 기준으로 순위를 계산하는 문서 검색단계; 및

검색된 단어와 및 검색용 태그를 포함하는 문장과 그 문장이 포함된 문서에 대한 내용을 표시하는 결과 표시단계;

를 포함하여 이루어지는 것을 특징으로 하는 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법.

**청구항 5.**
제 4항에 있어서,

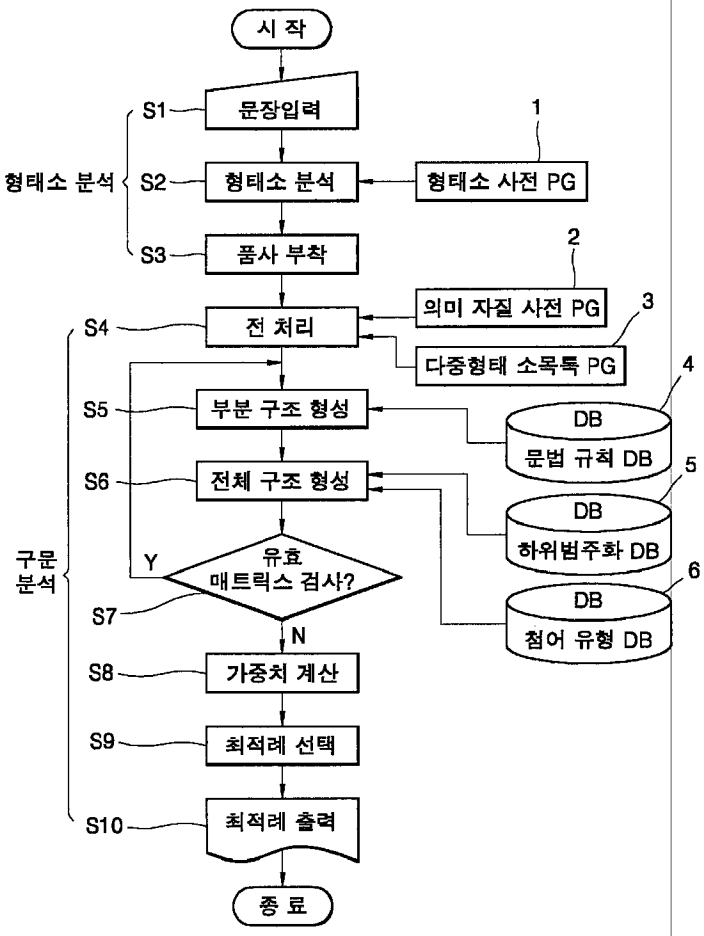
상기 문서 검색 단계는, 구문 분석된 정보만을 이용하여 질의어의 구문 분석 결과만을 기반으로 기 분석된 문서 데이터베이스를 검색하여 일치하는 내용을 추출, 제공하는 일반 검색 단계(모드)와, 질의어에 특정한 표현이 포함되는 경우, 검색자의 선택에 따라 특별검색 규칙정보와 명사체계 데이터베이스에 의해 특별 검색 모드를 위한 조건을 생성시키고, 이를 기반으로 의미적으로 주어진 성분에 종속된 내용들을 검색하여 제공하는 특별 검색 단계(모드)를 포함하고,

상기 일반 검색 단계는, 주어진 질의어의 직접 구성 성분과 일치하는 자료를 추출하여 제공하는 성분 일치 검색 방법과, 질의어를 구성하는 성분들을 포함하되 핵어인 술어와 의미적으로 유사한 술어들을 포함하는 자료를 추출하여 제공하는 의미 일치 검색 방법으로 이루어지며,

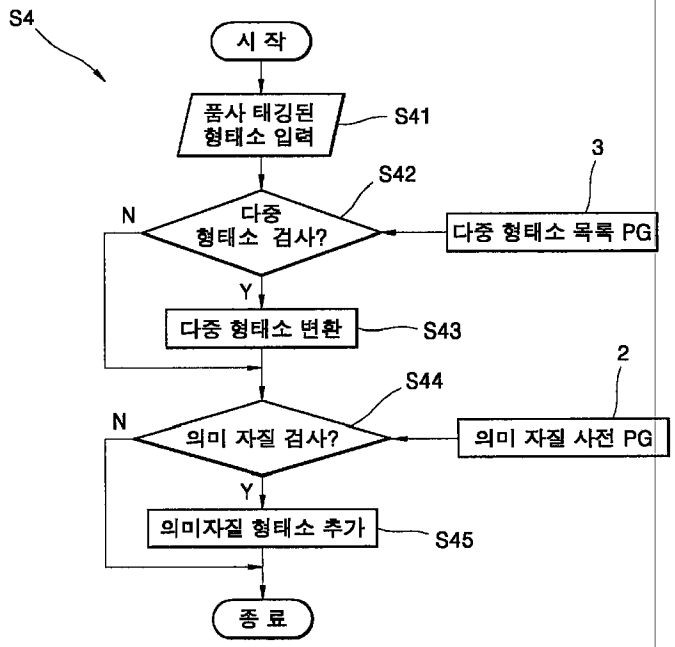
상기 특별 검색 단계는, 상기 특별검색 규칙정보와 명사체계 데이터베이스와 같은 명사의 의미적 위계 구조에 대한 데이터베이스를 이용하는 것을 특징으로 하는 모빌적 형상 개념을 기초로 한 구문 분석방법을 이용한 자연어 검색 방법.

도면

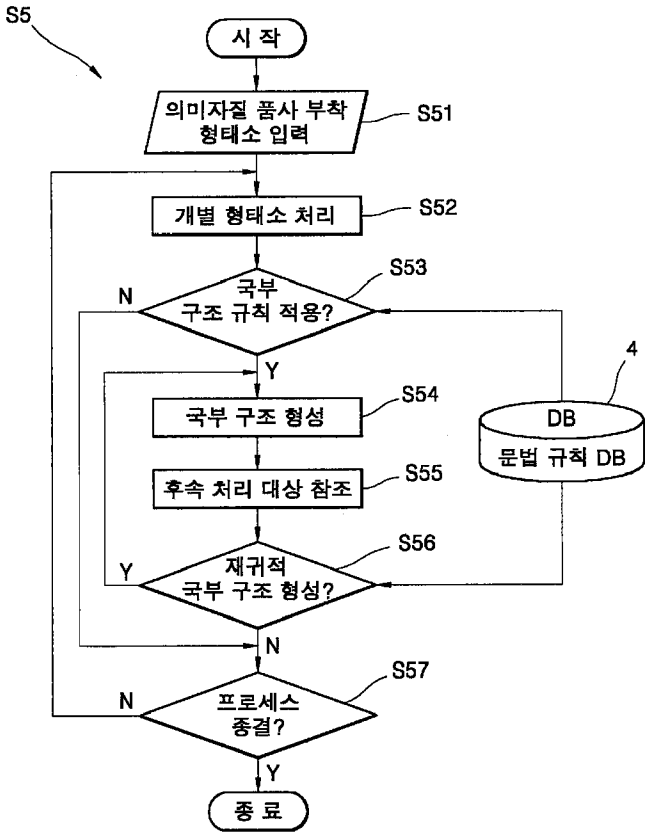
도면1


시 작
S1
문장입력
1
형태소 분석
S2
형태소 분석
형태소 사전 PG
S3
품사 부착
2
의미 자질 사전 PG
3
S4
전 처리
다중형태 소목록 PG
4
S5
부분 구조 형성
DB
문법 규칙 DB
5
S6
전체 구조 형성
DB
하위범주화 DB
6
구문 분석
Y
유효 매트릭스 검사?
S7
DB
첨어 유형 DB
N
S8
가중치 계산
S9
최적례 선택
S10
최적례 출력
종 료

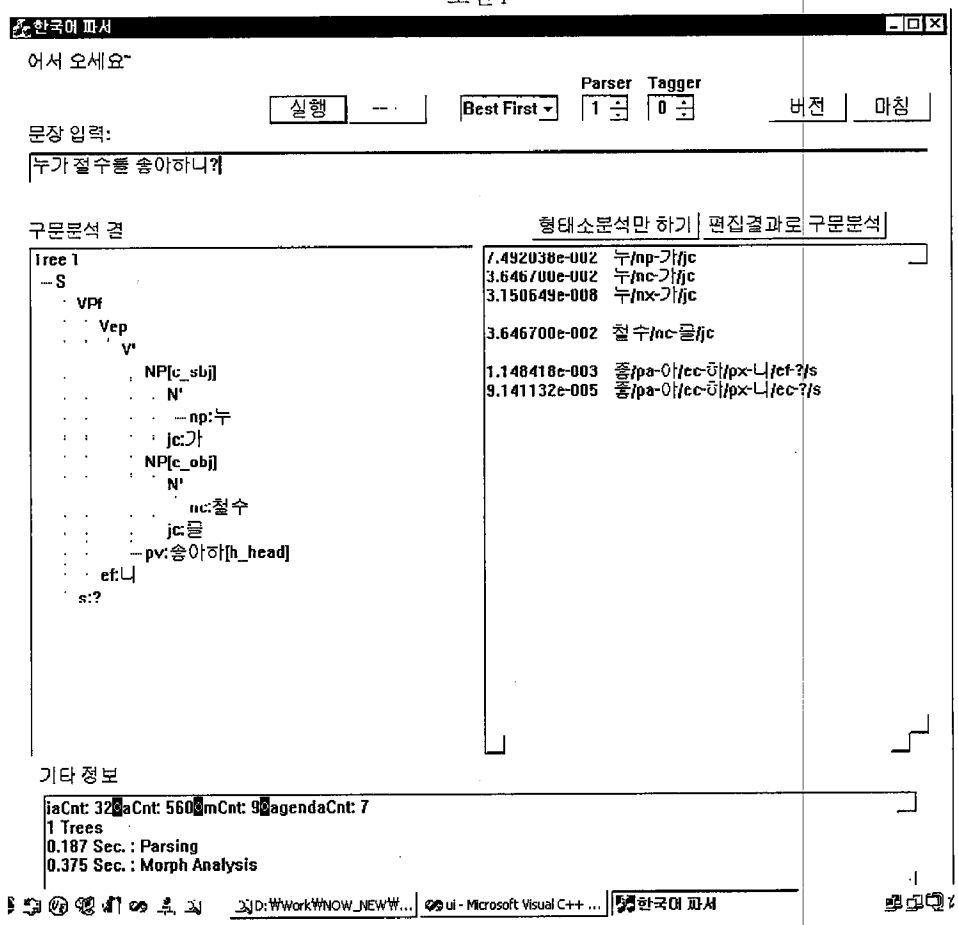
도면2


S4
시 작
품사 태깅된 형태소 입력
S41
3
S42
N
다중 형태소 검사?
다중 형태소 목록 PG
Y
다중 형태소 변환
S43
2
S44
N
의미 자질 검사?
의미 자질 사전 PG
Y
의미자질 형태소 추가
S45
종 료

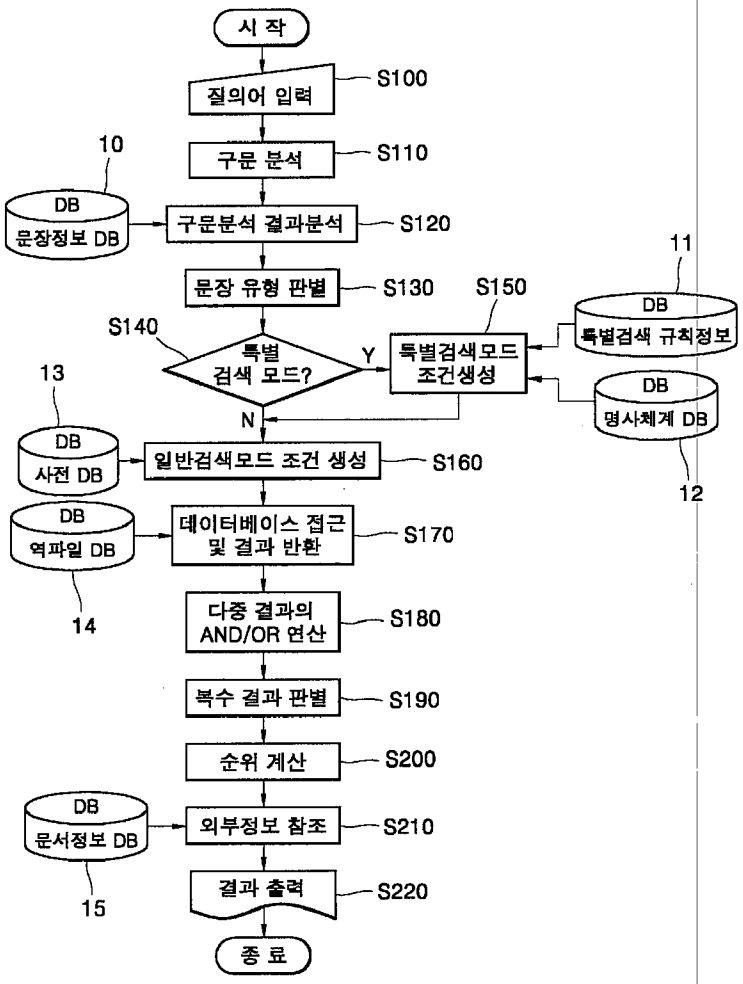
도면3


S5
시 작
의미자질 품사 부착 형태소 입력
S51
개별 형태소 처리
S52
국부 구조 규칙 적용?
S53
N
Y
국부 구조 형성
S54
후속 처리 대상 참조
S55
재귀적 국부 구조 형성?
S56
Y
N
프로세스 종결?
S57
N
Y
종 료
4
DB
문법 규칙 DB

도면4

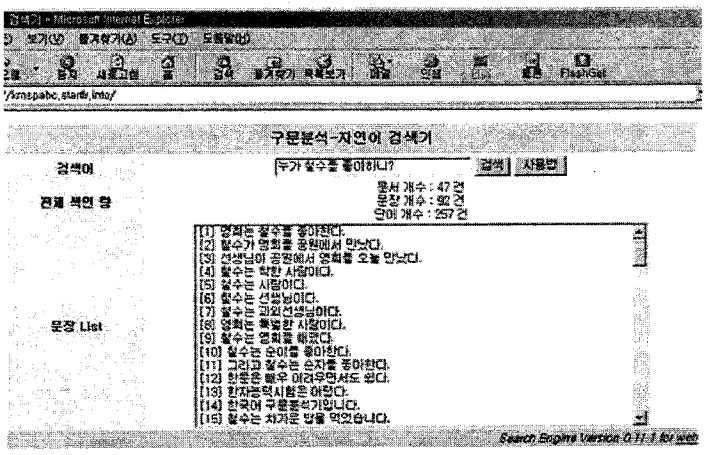

한국어 파서
어서 오세요~
실행
Best First
Parser
Tagger
1
0
버전
마침
문장 입력:
누가 철수를 송아하니?
구문분석 결
형태소분석만 하기
편집결과로 구문분석
Tree 1
S
VPf
Vep
V'
NP[c_sbj]
N'
np:누
jc:가
NP[c_obj]
N'
nc:철수
jc:를
pv:송아하[h_head]
ef:니
s:?
7.492038e-002 누/np-가/jc
3.646700e-002 누/nc-가/jc
3.150649e-008 누/nx-가/jc
3.646700e-002 철수/nc-를/jc
1.148418e-003 좋/pa-아/ec-하/px-니/ef-?/s
9.141132e-005 좋/pa-아/ec-하/px-니/ec-?/s
기타 정보
iaCnt: 32 aCnt: 560 mCnt: 9 agendaCnt: 7
1 Trees
0.187 Sec. : Parsing
0.375 Sec. : Morph Analysis
D:₩Work₩NOW_NEW₩...
ui - Microsoft Visual C++ ...
한국어 파서

도면5

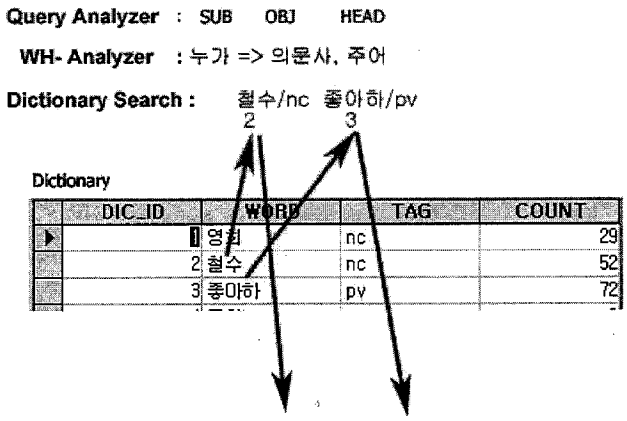



도면6

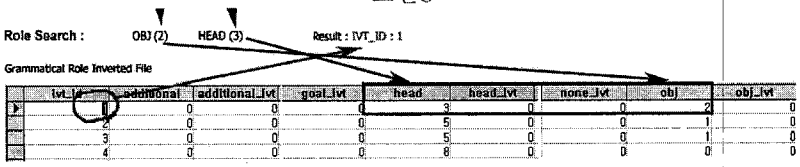
검색 질의어 : 누가 철수를 좋아하니?



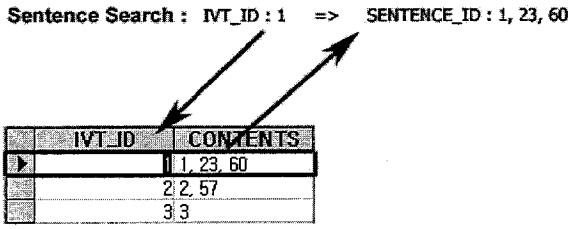
도면7



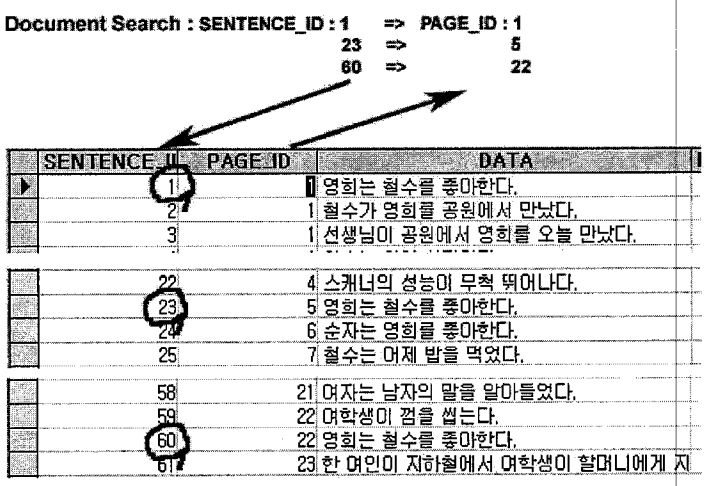

도면8




도면9




도면10



도면11

**Document Information (Filename, URL...) Search : PAGE_ID 1, 5, 22**

| | PAGE_ID | TITLE | LASTUPDATE | PAGEMAP |
|---|---|---|---|---|
| ▶ | 1 | ./data/a.txt | 1030986374 | 1,2,3,4,5,6,7,8 |
| | 2 | ./data/New01.txt | 1030986375 | 9 |
| | 3 | ./data/pjs01.txt | 1030986375 | 10,11,12,13,14,15,16,17,18 |
| | 4 | ./data/pjs02.txt | 1030986376 | 19,20,21,22 |
| | 5 | ./data/pjs03.txt | 1030986376 | 23 |
| | 6 | ./data/pjs04.txt | 1030986376 | 24 |
| | 7 | ./data/pjs05.txt | 1030986376 | 25,26,27,28 |
| | 20 | ./data/wsj01.txt | 1030986378 | 55,57 |
| | 21 | ./data/wsj08.txt | 1030986378 | 58 |
| | 22 | ./data/Wsj10.txt | 1030986378 | 59,60 |
| | 23 | ./data/Wsj10a.txt | 1030986379 | 61 |
| | 24 | ./data/Wsj10b.txt | 1030986379 | 62 |

도면12

**Print Result :**